

AN APPLICATION CONTAINER THAT ALLOWS CONCURRENT EXECUTION ON MULTIPLE NODES IN A CLUSTER

1 Technical Field

2 The present invention relates to a field of packages utilized on clusters of
3 computer systems. More specifically, the present invention relates to failing-over of
4 applications configured to run on clustered computer systems and a method of eliminating
5 the need to fail-over applications.

6 Background

7 Commonly, pluralities of computers, databases, printers and other computing or
8 computing related devices are often established in clusters or as members or elements of a
9 network, hereafter, collectively “Clusters”. Clusters are often defined as parallel or
10 distributed systems that consist of a collection of interrelated whole computers, that is
11 utilized as a single, unified computing source. As such, it is commonly appreciated that
12 Clusters commonly consist of computing devices (i.e., Nodes) and other peripheral
13 devices to which access thereto may be controlled by particular Nodes. Clusters enable
14 information system managers to share a computing load over several systems, thereby
15 enhancing the capabilities of the system as a whole, minimizing the adverse effects of
16 failures of certain Nodes or devices on a Cluster, and allowing for the expansion of
17 capacity and capabilities of a given Cluster by adding additional or different Nodes and/or
18 devices to the Cluster. Often Clusters are designed to be Highly Available (i.e., the
19 resources associated with the Cluster have a minimum operating reliability, often
20 measured in the 99.99% range and better). As is well known in the art, High Availability
21 implies that faults in services provided by a Cluster will be detected within a specified
22 time parameter and restored within a specified time parameter.

23 As Clusters have increased in use, size and complexity, products have been
24 developed that manage such Clusters. One such product is SERVICEGUARD®,
25 developed by HEWLETT PACKARD®. In order to provide the desired High
26 Availability, many of the Cluster management products today utilize a fail-over
27 mechanism when a fault is detected on a Node of a Cluster. The fail-over mechanism
28 basically provides that when a Node on which an application is running fails, for
29 whatever reason, the application is “failed-over” to another Node. In the fail-over
30 process, essentially, the application is restarted on a new Node within a given amount of
31 time. Cluster managers often utilize “Packages”, that contain all the resources that an

1 application might need in order to run on a Node. These Package essentially provide the
2 information needed by the fail-over Node in order to restart the application. Examples of
3 information a Package may contain include information relating to the type of storage an
4 application utilizes, the IP addresses that it uses for clients that connect to the application
5 and other information that basically enables the application to run on the system. Cluster
6 management tools, such as SERVICEGUARD®, often are capable of monitoring
7 Packages and starting them as necessary in order to activate an application on a Cluster.
8 Further, when a fail-over occurs, it is the Package that is usually provided to the fail-over
9 Node by the Cluster manager.

10 Further, in addition to providing a container of resources needed to fail-over an
11 application, Packages also provide other benefits. Since all the resources are in one
12 container, a Package makes it easy for Cluster managers to monitor and manage the
13 associated application. Cluster managers, via the Package, can see what components are
14 functioning optimally and also extend the application in interesting ways. For example, if
15 storage devices need to be added for use with a given application, the needed devices can
16 merely be added to the Package and then be available to the application regardless of the
17 Node on which the Package is currently running.

18 However, as Clusters become more common, many new applications are being
19 developed which are essentially, “Cluster Aware”, i.e., they appreciate the fact that the
20 application is not merely running on a single computer and instead is running on a Node
21 of a Cluster. With Cluster Aware applications, it is often undesirable to require an
22 application to fail-over to a new Node whenever a fault is detected, because such fail-over
23 is often inefficient, and needed data is often lost. Further, many of these Cluster Aware
24 applications desire to implement instantiations of themselves on every Node in the
25 Cluster. As such, they often do not behave as non-Cluster aware applications when a fail-
26 over occurs because of the instantiations of themselves often are not amenable to the
27 commonly utilized fail-over processes. Examples of Cluster aware applications include
28 volume management services, and ORACLE® database applications, where the database
29 is implemented on multiple Nodes for performance purposes.

30 In particular, ORACLE® database operations are commonly implemented on a
31 Cluster such that each Node has a specific instance of the ORACLE® application. Each
32 Node is capable of implementing the application but does not know that the application
33 also exists on other Nodes of the Cluster. As such, the application is commonly
34 implemented simultaneously on multiple Nodes, thereby wasting and inefficiently

1 utilizing Cluster resources. Thus, a need exists for a process which enables Cluster aware
2 applications to maintain High Availability without requiring the application to fail-over to
3 other Nodes.

4 One way to provide for the multiple instantiations of applications on multiple
5 Nodes in a Cluster manager system is to hard-code the multiple instantiations into the
6 source code for the Cluster management software. This hard-coding enables the Cluster
7 management software to appreciate, at start-up, that multiple copies of an application are
8 to be implemented on the Nodes of the Cluster and to automatically start up the multiple
9 instantiations of the applications on the multiple Nodes. While this process works for
10 existing applications (whose needs are known), it is not very adaptable since any new
11 application desiring to be Cluster Aware requires recoding of the source code for the
12 Cluster management software. Thus, there is a need for a process that provides High
13 Availability to Cluster Aware applications while being agnostic as to the particular
14 application being implemented.

15 **Summary**

16 The present invention provides a process for providing High Availability in a
17 Cluster without utilizing a fail-over mechanism. More specifically, the present invention
18 utilizes a Multi-Node Package (MNP) to concurrently run on multiple Nodes on a Cluster
19 and thereby provide the desired operational performance without the downtime associated
20 with fail-over operations.

21 Instead of initiating a Package on a single Node, the present invention initiates a
22 Package on every Node (or a desired percentage of Nodes) on a Cluster. Instead of an
23 application failing-over to another Node, to ensure High Availability, the MNP merely
24 shifts responsibilities for an application having faults or even possible faults on a first
25 Node to another Node at which another instance of the Package is operating.

26 **Description of the Drawings**

27 The detailed description will refer to the following drawings, wherein like
28 numbers refer to like elements, and wherein:

29 Figures 1A and 1B are representative flow charts of how a Cluster management
30 system implementing the present invention utilizes the MNP of the present invention to
31 provide High Availability applications.

32 **Detailed Description**

33 As mentioned previously, the present invention provides a process for providing
34 High Availability of applications implemented on a Cluster by utilizing MNPs which do

1 not need or utilize a fail-over mechanism. The process utilizes any Package well known
2 in the art and does not require specialized Packages or other applications. One example
3 of a Package description is provided below:

```
4 ****
5 ***
6 **** HIGH AVAILABILITY PACKAGE CONFIGURATION FILE (template)
7 ***
8 ****
9 ***
10 ***** Note: This file MUST be edited before it can be used.
11 ****
12 *** For complete details about package parameters and how to set them,
13 ****
14 *consult the MC/SERVICEGUARD or SERVICEGUARD OPS Edition
15 manpages**
16 *** or manuals.
17 ****
18 ****
19 ***
20 # Enter a name for this package. This name will be used to identify the package
21 when viewing or manipulating it. It must be different from the other configured
22 package names.
23 PACKAGE_NAME          pencil
24 PACKAGE_TYPE          MULTI_NODE
25 NODE_NAME              *
26 RUN_SCRIPT             /etc/cmcluster/pencil/pencil.control
27 RUN_SCRIPT_TIMEOUT     NO_TIMEOUT
28 HALT_SCRIPT            /etc/cmcluster/pencil/pencil.control
29 HALT_SCRIPT_TIMEOUT    NO_TIMEOUT
30 SERVICE_NAME           samba_nmb_pencil
31 SERVICE_FAIL_FAST_ENABLED NO
32 SERVICE_HALT_TIMEOUT   300
33 SERVICE_NAME           samba_smb_pencil
34 SERVICE_FAIL_FAST_ENABLED NO
35 SERVICE_HALT_TIMEOUT   300
36 SUBNET                 15.13.168.0
37 AUTO_RUN                YES
38 LOCAL_LAN_FAILOVER_ALLOWED YES
39 NODE_FAIL_FAST_ENABLED NO
```

40
41 By utilizing commonly available Packages, the present invention provides
42 programmers and system designers with tremendous latitude and adaptability, as
43 specialized code is not needed nor are changes to the source code required in order to
44 expand an application's capabilities or to implement a Cluster aware application on a
45 Cluster.

1 The present invention is configured such that when the Cluster management
2 software detects a new Cluster aware application that is to be implemented on the Cluster,
3 the software obtains the Package associated with one Node of the Cluster, copies the
4 Package and loads the Package onto the desired other Nodes of the Cluster. The Cluster
5 management software knows of the existence of the Package on the other Nodes of the
6 Cluster, such that whenever a fault occurs on one Node, other Nodes may be utilized to
7 immediately implement the application without requiring a fail-over routine copying the
8 needed information contained in the Package.

9 For one embodiment of the present invention, whenever an application is initiated
10 on one Node of the Cluster, it also must have a Package initiated on every other Node of
11 the Cluster. This embodiment is referred to as the System MNP (SMNP) and is
12 commonly utilized when an application utilizes a Cluster membership roster to
13 communicate amongst themselves.

14 However, another embodiment of the present invention does not require every
15 Node to be running an instance of the Package. This configuration is referred to as the
16 General Purpose MNP (GMNP) and might be utilized to balance a load on a Cluster
17 while also providing High Availability. For example, an application might be configured
18 to be run on only 50% of the Nodes, instead of all the Nodes on the Cluster. When a fault
19 occurs on one Node, a second Node (for example, from the 50% containing the Package)
20 suitably continues the operation of the application without requiring a fail-over to occur.

21 Referring now to Figures 1A and 1B, one embodiment of how a system utilizing
22 the MNP of the present invention might be configured to operate in order to provide High
23 Availability to an application running on a Cluster is shown. As shown, this process
24 begins when an application is to be loaded onto a Node of a Cluster (Block 102). At the
25 time of loading, the Package associated with the application is configured into the Cluster
26 by informing the Cluster manager that there is a configuration change. This notification
27 may be accomplished through a command, an input via a Graphical User Interface (GUI),
28 or other methods known in the art. At this time, the Cluster manager performs some
29 verifications of the new Package by confirming that all the attributes specified in the
30 Package configuration are legitimate. One of these attributes will indicate whether the
31 Package is a failover type, a GMNP type, or a SMNP type.

32 If the Package is a failover type (Block 104) the processing ends and the Cluster
33 manager knows that if a failover condition occurs, it will restart the Package on another
34 Node (Block 106). The processing then continues with Block 110, as explained

1 hereinbelow. If the application is an MNP application (i.e., either a GMNP, and SMNP,
2 or any other type of MNP), the failover behavior will not occur, but instead an instance of
3 the Package will be started on all the Node (in the case of SMNP) or some subset of the
4 Node (in the case of GMNP) (Block 108).

5 Further, it is to be appreciated that an application may or may not be Cluster
6 Aware and still may utilize an MNP. However, since a Cluster Aware application
7 generally knows that it has peer instances running on other Node in the Cluster, the
8 present invention makes it easier to configure Cluster Aware applications so that how the
9 specific application behaves (during start-up and during the handling of failures by
10 instances of the application) is better controlled by the Cluster manager.

11 Referring again to Figures 1A and 1B, Block 108, at this instance all Nodes, or
12 those designated Nodes, on the Cluster contain an instantiation of the Package (when the
13 Package is of the MNP type). It is to be appreciated, that for certain applications, the
14 Package may be an extremely large segment of operating instructions, software codes and
15 other information that may be implemented by each Node on the Cluster. As such, the
16 present invention may increase the load requirements for the Cluster as a whole, while
17 eliminating the need to fail-over applications. Thus, those skilled in the art appreciate
18 that, in certain circumstances, the General Purpose MNP embodiment may consider
19 trade-offs between load factor and fail-over success that occur with the implementation of
20 the present invention in deciding upon which Nodes a Package is to instantiated.

21 Referring again to Figures 1A and 1B, after the Package has been loaded on the
22 desired Nodes of the Cluster, processing then continues with the application itself being
23 implemented on the desired Node, i.e., herein the primary Node (Block 110). The
24 application then continues to process on the primary Node as according to its normal
25 routines and specifications. Meanwhile the Cluster management software preferably
26 continues to monitor the implementation of the application and seeks to detect faults in
27 the application (Block 112).

28 If a fault in the application is not detected, the application continues processing on
29 the primary Node (Block 114) until either the application is terminated (Block 116) or a
30 fault is detected (Block 112), whichever occurs first.

31 When a fault is detected, the Cluster management system suitably proceed along
32 one of two paths (Block 113). If the Package is a MNP, the Cluster management takes
33 appropriate actions which may include instructing a second Node (at which an
34 instantiation of the Package was previously loaded) to assume the processing

1 responsibilities for the application (Block 115) or even to stop attempting to communicate
2 with the, now faulted, primary Node. It is to be appreciated that various faults may occur
3 in an application. As such, various responses to such fault may be directed by the Cluster
4 manager. Both the faults and possible responses thereto are too numerous to list here. As
5 such, it is to be appreciated that the Cluster manager takes appropriate actions, which may
6 include notifying other Nodes of the failure (when the application is Cluster aware and
7 such notifications are appropriate), ceasing operations on the faulted Node, shifting
8 application processing to another Node, and other actions. Further, when the application
9 is Cluster aware and multiple instantiations of the application are being provided on at
10 least the second Node, the transfer of processing capabilities can occur without a fail-over
11 and preferably without any interruption in the data processing flow.

12 Once, if at all, the processing has been transferred to the second Node, the Cluster
13 management system may then attempt to determine what caused the error on the primary
14 Node and, if possible, correct the error. In certain embodiments, upon the correction of
15 the error on the primary Node, the responsibility for implementing the application may be
16 transferred back to the primary Node. Such a transfer might be desirable for load
17 balancing purposes, proximity to data files, proximity of the Node to clients and/or for
18 similar and other reasons. However, the process may also be implemented such that the
19 second Node continues to process the application for an indefinite time period.

20 Referring again to Block 113, when the Package is not an MNP package, the
21 processing then proceeds to Block 117. At this instance, the application is preferably
22 failed-over to a second Node and implementation of the application is resumed thereon.

23 The process then resumes with looping through Blocks 112-114-116 until either
24 the application is terminated or another fault occurs. It is to be appreciated that upon the
25 second Node assuming the responsibilities for implementing the application, the second
26 Node effectively becomes the primary Node and that the processing responsibilities may
27 be transferred to a third Node on the Cluster, when available, as the detection of faults or
28 other operating concerns dictate.

29 Further, it is to be appreciated that not every Package is capable of providing
30 notification of instance failure. In certain embodiments, Package status and Package
31 status change notifications may be provided via Application Program Interfaces (APIs)
32 and other well known devices. While such notifications are preferred, the present
33 invention does not depend upon such notifications for its operation. Further, with MNPs
34 it is to be appreciated that such APIs may also be configured to let a Cluster manager

1 know that an instance of the package has changed state on a particular Node (i.e., started
2 or stopped). Such capabilities may be desired for Cluster aware applications that
3 communicate among their peer instances and may desire to transfer their load from a
4 down instance to an up instance.

5 Further, it is to be appreciated that the process flow shown in Figures 1A and 1B,
6 for purposes of simplicity, combines configuration and instance fault detection into one
7 process flow. Each of which may be considered to be a separate process (i.e., the
8 configuration steps may be considered distinct of the fault detection steps, and vice
9 versa). More specifically, configuration usually occurs without any instances of the
10 applications running. As such, it is not a common thing to make a non-Cluster aware
11 application Clusteraware. The Cluster-aware application is either configured as a set of
12 disjoint traditional failover packages (one package for each Node in the Cluster and these
13 Packages do not specify failover nodes) or it is configured as a MNP.

14 As eluded to earlier, the MNP of the present invention provides certain advantages
15 over the failover type of Package. First, with the MNP, Cluster management is more
16 efficient. If the Cluster manager doesn't provide an MNP, the application is kind of
17 "shoe-horned" into disjointed failover Packages. When these disjointed Packages fail, in
18 order to halt the application on all the Nodes, the Cluster manager has to individually
19 issue a halt command for each Package.

20 Second, Cluster management issues arise when a new node is added to the
21 Cluster. For a SMP in particular, one generally wants to verify that when an existing
22 SMP is configured on the Cluster and the Administrator wants to configure a new Node
23 into the Cluster, that SMP will be able to start an instance of itself on that new node. This
24 verification is built into the "All-Node Package", which might be a SMP that concurrently
25 runs on all up Nodes or it might be a traditional failover package that can failover to any
26 Node in the Cluster.

27 Further, once a Cluster manager decides whether they want a MNP or a traditional
28 failover package and configure the application Package as such, the Cluster manager then
29 knows to take a different action based upon the type of application Package. If it is a
30 traditional failover package it will failover the application to an adoptive Node. If it is a
31 SMP, it notes the failure and, if appropriate, notifies any Cluster-aware Nodes that may
32 have registered for such notifications of the failure. Further, if the Package is a GMNP, it
33 will do the same thing as a SMP, but in a "load balancing" embodiment. When load
34 balancing, the Cluster manager utilizes the specifications for the Package to determine

1 how to transfer the application. For example, such load balancing may indicate that that
2 Package should run one at least “x” number of Nodes or at least “y”% of the Nodes in the
3 Cluster. As such, the cluster manager may try to start another instance of the package on
4 a Node where an instance of that package is not currently running, or, if the load balance
5 thresholds are being met, the Cluster manager may not start another instance of the
6 Package. A starting of another instance of the Package on another Node, however, is not
7 a failover because any running instances on Nodes that did not fail will continue to run
8 while the new instance is started on the new Node.

9 As such, the present invention provides for High Availability on Cluster aware
10 applications, without using a fail-over technique, by utilizing multiple instantiations of
11 Packages on multiple (if not all) Nodes of a Cluster. Such MNPs provide for near
12 seamless transitions of applications across Nodes.

13 Further, it is also to be appreciated that the processes of the present invention may
14 be suitably utilized to also determine system utilization levels, application efficiencies
15 and other information relating to an implementation of an application.

16 Lastly, while the present invention has been described in the context of process
17 flows, Packages, Clusters and networked environments, it is to be appreciated that the
18 present invention is not so limited. Containers of information (whether they be Packages
19 or in other formats) may be utilized in accordance with the present invention for other
20 and/or additional purposes as provided for in the detailed description, drawing figures and
21 claims.